Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models

Year: 2023 | Citations: 418 | Authors: Rongjie Huang, Jia-Bin Huang, Dongchao Yang, Yi Ren, Luping Liu

---

Abstract

Large-scale multimodal generative modeling has created milestones in text-to-image and text-to-video generation. Its application to audio still lags behind for two main reasons: the lack of large-scale datasets with high-quality text-audio pairs, and the complexity of modeling long continuous audio data. In this work, we propose Make-An-Audio with a prompt-enhanced diffusion model that addresses these gaps by 1) introducing pseudo prompt enhancement with a distill-then-reprogram approach, it alleviates data scarcity with orders of magnitude concept compositions by using language-free audios; 2) leveraging spectrogram autoencoder to predict the self-supervised audio representation instead of waveforms. Together with robust contrastive language-audio pretraining (CLAP) representations, Make-An-Audio achieves state-of-the-art results in both objective and subjective benchmark evaluation. Moreover, we present its controllability and generalization for X-to-Audio with"No Modality Left Behind", for the first time unlocking the ability to generate high-definition, high-fidelity audios given a user-defined modality input. Audio samples are available at https://Text-to-Audio.github.io